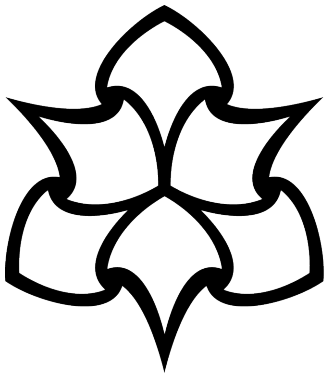


The Use of Control Charts by Lay-people and Hospital decision-makers for Guiding Decision Making

Word count: 7820

Schmidtke, K. A.,[1] Watson, D. G.,[2] Vlaev, I.[3]

[1] University of Warwick, Behavioural Science Group; Kelly.Schmidtke@wbs.ac.uk
[2] University of Warwick, Psychology, D.G.Watson@warwick.ac.uk
[3] University of Warwick, Behavioural Science Group, Ivo.Vlaev@wbs.ac.uk

*Corresponding author is Kelly Ann Schmidtke

Behavioural Science Group
Warwick Business School
The University of Warwick
Coventry, CV4 7AL, UK,
e-mail: Kelly.Schmidtke@wbs.ac.uk.
Phone: 07758933026

Acknowledgement

This article presents independent research commissioned by the National Institute for Health Research (NIHR) under the Collaborations for Leadership in Applied Health Research and Care (CLAHRC) programme, West Midlands. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. We also thank Dr Richard Lilford for his encouragement throughout this project.

Abstract

Graphs presenting healthcare data are increasingly available to support lay-people and hospital staffs' decision-making. When making these decisions, hospital staff should consider the role of chance, i.e., random variation. Given random variation, decision-makers must distinguish signals (sometimes called special-cause data) from noise (common-cause data). Unfortunately, many graphs do not facilitate the statistical reasoning necessary to make such distinctions. Control charts are a less commonly used type of graph that support statistical thinking by including reference lines that separate data more likely to be signals from those more likely to be noise. The current work demonstrates for whom (lay-people and hospital staff) and when (treatment and investigative decisions) control charts strengthen data driven decision-making. We present two experiments that compare people's use of control and non-control charts to make decisions between hospitals (Funnel charts vs League tables) and to monitor changes across time (Run charts with control lines vs Run charts without control lines). As expected, participants more accurately identified the outlying data using a control chart than a non-control chart, but their ability to then apply that information to more complicated questions (e.g., where should I go for treatment?, and should I investigate?) was limited. The discussion highlights some common concerns about using control charts in hospital settings.

Keywords: statistics, decision-making, healthcare

The Use of Control Charts by Lay-people and Hospital decision-makers for Guiding Decision Making

Graphs containing healthcare information are increasingly made available to support data driven decision-making (Wainer, 2013). Unfortunately, such graphs commonly do not facilitate the statistical reasoning necessary for such decisions (Chance, 2002). The present study examines both lay-people's (i.e., university students) use and hospital staff (people positioned to make decisions based on safety and quality performance measures) use of a type of graph designed to support statistically informed decision-making, specifically control charts.

Control charts support statistical reasoning.

Control charts support statistically informed decision-making by including reference lines that highlight the role of chance (Shewart, 1939; Woodall, 2006).[1] Control charts typically include a centre line indicating the central tendency of data and at least two control lines signifying chance variation (Thor, et al., 2007). Examples of such charts are provided in the top half of Figure 1.

The data falling between the upper and lower control lines are likely due to common-cause variation. Common-cause variation are naturally anticipated fluctuations in any working process, i.e., chance. Note that by 'chance' we do not mean that the data has no cause, its cause is likely something that can be expected within the working process. If any common-cause data are unacceptable then something about the process that underlies all the common-cause data needs to be adjusted for total quality improvement. In contrast,

[1] One may note that hospital data are often presented as tables and that people often prefer tables over graphs (Hildon, Allwood & Black, 2012). Past research comparing people's use of tables and graphs has found that the two presentation methods are best suited to answer different types of questions (Speier, 2006; Vessey I. 1991). While tables help decision-makers better identify past, unique data, graphs are better at portraying patterns in the data, e.g., 'In what month was the percent of patients waiting more than 4 hours to be seen in A&E highest?' Versus 'Is this percent increasing?' As quality improvement relies on recognizing patterns in data, we concentrate on graphs in the current work.

data falling outside the control lines are more likely to be the result of a special-cause, which an investigation could discover (Deming, 1979). Logically any data point that is not due to common cause must be due to special cause. However, in practice all control charts reveal is the likelihood that data are one or the other and so guide when further investigative action is most likely to be useful. For graphs presenting between-groups comparisons, special-cause data indicate any group performing better or worse than expected by chance, see Figure 1A. For graphs presenting time-series comparisons special-cause data indicate a time when performance was better or worse than expected by chance, see Figure 1B.

For proportion data, the width of the upper and lower control lines can be adjusted for unequal sample-sizes point by point. When between-group comparisons are considered, as in Figure 1A, the adjusted control lines take on a funnel-like appearance by arranging the groups on the horizontal axis so that the group representing the smallest sample-size appears first followed by those representing larger sample-sizes. Such charts are called funnel charts (Spiegelhalter, 2005). When time-series comparisons are considered, as in a run chart, the adjusted control lines take on a step-like appearance. The steps' extremity depends on the diversity of the sample-sizes; as the sample-sizes in Figure 1B are similar the steps' extremity are slight (See Polit & Chaboyer, 2012, Figure 3 for an example of control charts with more extreme steps).

Determining where the control lines are set on charts for different measures should reflect the cost of investigating and the cost of not investigating, in terms of financial cost, quality, and harm. The control lines in Figure 1 are set at 3 standard deviations from the mean. More cautious decision-makers may think this is too lenient and prefer a two standard deviation control line. Such adjustments are not without consequence, as this will increase the chance of a false positive up to 25% depending on the underlying distribution (Kvanli, et al. 2006). Thus without

knowledge of the underlying distribution, as is common in healthcare, many prefer to set the control lines at 3 standard deviations. Further instruction on control charts' construction and interpretation is outside the scope of the current paper; for further details please see Amin, 2001 or Muhammad, Worthington & Woodall, 2008.

Common graphical methods that do not support statistical reasoning.

League tables and run charts without control lines are more commonly used types of graphs, both of which often do not highlight the role of chance. Examples of such charts are provided in the bottom half of Figure 1. League tables present between-groups comparisons by listing each group's performance in rank-order, see Figure 1C. Such tables are used widely by healthcare regulators and may be used to inform patients (Peymané, et al, 2002). While the ordering in league tables renders the best and worst performers clear, it does not highlight chance variation (Goldstein & Spiegelhalter, 1996). As a result, people using such tables may deem the worst performing hospital as uniquely bad when it is in fact within chance. This is especially a problem when the groups contain different sample-sizes (Tversky & Kahneman, 1974). As discussed previously, funnel charts mitigate this problem by surrounding each group with statistical limits based on sample-size. Lay-people's use of funnel charts and league tables are compared in Experiment 1.

Run charts present time-series comparisons, see Figure 1D. Analyses of run charts are often guided by four basic rules to determine whether a process is unstable (Perla, Provost & Murray, 2011):

- shift - six or more consecutive points either all above or all below the median
- trend - five or more consecutive points all going up or all going down

- runs- too few or too many runs or crossings of the median line, with 'too many' given according to tabled critical values one must look up
- astronomical point – a point that is different from the rest of the points

Notable, precise definitions are available for the first three rules, but the last rule depends on chance variation, which is often not obvious in run charts without control lines.

Sometimes people mistakenly identify common-cause data as special-cause data, i.e., false positives. This is true for both between-groups and time series analyses (Marshall, Mohammed, & Rouse, 2004; Speekenbrink, Twyman & Harvey, 2012). This is a problem because investigations of unique data points within chance variation are unlikely to find special causes, and could divert resources from investigating/altering the entire process. The addition of control lines to run charts mitigates this problem by including statistical reference lines that contextualise all data as falling within or outside of chance variation.[2] People's use of run charts with and without control lines are compared in Experiments 1 (lay-people) and Experiment 2 (hospital staff).

Can people use control charts effectively?

The current paper compares people's use of between-group and time-series control charts to commonly used non-control charts. Below we briefly review recent research that has explored the use of control charts.

[2] For simplicity, the current study focuses on a single type of control chart and a single rule Western Electronic rule to identify irregular data, i.e., any data point outside 3SD is irregular. More complicated run charts that include multiple sets of control lines (e.g., one set at 2 SD and another set at 3 SD's) enhance decision-makers' ability to identify lower level statistically irregular trends. For example, while one data point outside of three SD is statistically irregular, two consecutive data outside 2 SD are irregular. While it is tempting to apply as many rules as possible, decision-makers ought to remain cautious as the more rules applied the greater the probability of a false positive. For more information see Amin, 2001.

Between-groups comparisons. More and more, patients are being allowed to choose which hospital they would like to be treated at for non-emergency procedures. To appreciate how lay-people make this choice, Hildon, Allwood and Black (2012) asked people to examine different displays of hospitals' performance measures to decide which hospital they would prefer for treatment. People had more difficulty understanding, and largely did not prefer funnel charts compared to other presentation methods (e.g., bar charts). However, after researchers explained the purpose of and how to interpret funnel charts, more numerate people warmed to them. But this should not be taken so far as to suggest that people can interpret them properly without assistance (Zikmund-Fisher, Smith, Ubel & Fagerlin, 2007).

Hospital decision-makers' use of funnel charts and league tables has already been compared in a randomized controlled trial (Marshall, et al., 2004). In that paper, board members were presented with either three funnel charts or three league tables displaying health service providers' 30-day mortality rates. After examining each graph board members were asked, if "they would take action as a result of the data, and if so to identify the service providers towards whom action would be directed" (p. 310). Those board members who received funnel charts recommended significantly fewer investigations (*M*s = 0.5, 1.0, 0.2) than those board members who received league tables (*M*s = 0.9, 4.5, 1.6). **Thus, randomized controlled trial evidence exists to say that hospital decision-makers' calls for investigative action are affected (in this case restrained) by presenting the data in a funnel chart rather than a league table.**

More recently, Rakow, Wright, Spiegelhalter and Bull (2014) examined lay-people's interpretation of funnel charts displaying different hospitals' mortality or survival rates. To determine if people understood the basic information presented in funnel charts, they were first asked questions that did not require them to consider the role of chance (e.g., which hospital has

the highest survival rate). After this people were asked to imagine they were going for treatment and to say which of two designated hospitals they preferred. Their preferences were sensitive to sample-size, suggesting that funnel charts might help people appreciate the role of chance.

While Rakow et al.'s findings are encouraging and inform the current study, two limitations should be noted. First, the two designated hospitals people were asked to choose between both often fell within the control lines. Therefore there was usually no statistically outlying reason to prefer either hospital. The present experiments addresses this concern by using hypothetical data, so that the two designated hospitals have equal performance measures but due to different sample-sizes one data point falls within and the other outside of the control lines. The second limitation is that Rakow's study did not assess participants' use of non-control charts, and so inferences that funnel charts facilitate statistical reasoning better than non-control chart methods might be premature. The current study addresses this second concern by assessing lay-people's use of funnel charts compared to league tables.

Time-series comparisons. The current study also compares people's use of run charts with and without control lines. Previous work suggests that run charts with control lines can improve healthcare providers' management of many variables, such as asthma attacks, infections, medical errors, and so on. (Alemi & Neuhauser, 2004; Carey & Teeters, 1995; Curran, Benneyan, & Hood, 2002). However, a limitation of many such studies is that they are often repeated measures designs without control groups.

At least one randomised controlled field trial has been performed by Curran, et al. (2008). Different hospital wards were or were not provided with control charts displaying their infection performance over several months. Those wards that received control charts decreased their infection rates more than those that did not, but not significantly so. Plausibly, significant

differences were not obtained because staff in wards that received control charts interacted with staff in wards that did not, thereby bolstering the latter's performance. The present experiments follow this work but use the alternative approach of a highly controlled laboratory design.

Hypotheses.

The design of control charts causes statistically irregular data to pop-out (e.g., Tresiman & Gelade, 1980). Therefore, control charts should empower even people untrained in control charts use to identify special-cause data. Accordingly our first hypothesis is that control charts have a strong advantage over non-control charts:

H1: People are more likely to accurately identify a special-cause datum when provided with a control chart than a non-control chart.

In contrast, control charts do not immediately tell people how to make more difficult decisions, which require further inferences and the applications of further principles. Therefore control charts are likely less able to help people untrained in control chart use to make treatment or investigative choices based on solid statistical reasoning. Regarding treatment choices, lay-people are asked to choose which hospital is more likely to see them within two weeks, may substitute **the intended statistical question with largely non-statistical questions. Specifically, instead of focusing on the variability presented between-hospitals, these non-statistical questions, will likely focus on whether the hospital is 'small' or 'large', for example: *Would I feel more comfortable in small or large hospitals?*, or *Based on my previous beliefs, are small or large hospitals faster?*** (Kahneman, 2003). Regarding investigative choices, hospital staff asked to monitor data over time may choose to rely on intuitive judgment to determine when investigations are warranted instead of statistical reasoning. Thus, our second hypothesis is as follows:

H2: People's responses to more difficult choices (i.e., which hospital to be treated at or whether to investigate) may not differ when provided with a control chart than with a non-control chart.

Experiment 1. Lay-people

Methods

We compare lay-people's use of control and non-control charts with regard to two types of decisions: deciding between different hospitals (i.e., the between-hospitals comparisons), and monitoring performance over time (i.e., the time-series comparisons).

Participants. One-hundred and seventy participants from the University of Warwick completed an on-line survey (59% Female, M_{age} = 21.3 years, SD = 3.5). They were from the Science (N = 57), Business (N = 55) Social Science (N = 46), and Arts Schools (N = 8), or did not say which school (N = 4). No participants said they were part of the Medical School. The majority of these participants (66%) had taken at least one statistics course, but few (16%) knew of control charts. Following completion of the survey, participants were entered into a lottery draw for a chance to win an Amazon gift voucher.

Experimental design. The experiment compared the responses of participants who examined a control chart with those who examined a non-control chart. Each participant sequentially examined and answered a set of questions about two randomly selected graphs. The first set of questions asked them to examine either a funnel chart or a league table (a between-hospitals comparison), and the second set of questions asked them to examine a run chart that either had or did not have control lines (a time-series comparison).

Materials. The survey was created and delivered online using *Qualtrics*©2012. Hypothetical data were used to create the charts in Excel.

Graphs presenting between-hospitals comparisons. The four graphs presenting between-hospitals comparisons contained data representing 15 hospitals' compliance with the maximum two week wait from receipt of an urgent GP referral for suspected cancer to the date patients are first seen. The original data set was transformed, flipped around its **grand proportion maintaining whole numbers,** to create a second data set and both were presented as funnel charts and league tables.[3] For an example of such transformation see Appendix A which contains all the graphs. Creating exact transformation of the data was not possible because of the grand proportion (i.e., the centre line) changes when the data are flipped. Within each data set, two hospitals had equal performance (i.e., 99% or 91%) obtained from different sample-sizes (i.e., 201 vs 623). Figure 1A presents one of the funnel charts, and Figure 2A presents its complimentary league table.

Graphs presenting time-series comparisons. The eight graphs presenting time-series comparisons contained data representing the percentage of patients waiting more than four hours to be seen in a single hospital's emergency department across 12 months. Within each graph the last datum was placed either within or outside of the control line. The original two data sets (one with the last data point within and the other outside of the control lines) were transformed, **flipped around its grand proportion maintaining whole numbers**, to create a second data set,

[3] As we wanted the percentages displayed in the charts to convey something that could actually happen, and so whole numbers were needed for the numerator of each hospital's performance. To do this we first found the flipped percentages. From the original data set, we deducted each hospital's percent compliance from its grand mean. Then we deducted these differences from the grand mean. Then, second, to find the new numerator, we multiplied each hospital's denominator by its flipped percentage, and rounded to the nearest whole number. The second funnel chart contains each hospital's percent compliance based on the rounded numerator and the original denominator.

and both were presented as run charts with and without control lines.[4] An exact transformation was not possible, because whether the data is in- or out-of-control depends on the average value which was different for increasing and decreasing data; so adjustments were made to preserve the in- and out-of-control states of the last datum. Figure 1B presents one of the run charts with control lines, and Figure 2B shows the complimentary run chart without control lines. An array of these time-series graphs can be viewed in appendix B.

Procedure. **The survey was set up to randomly present the selected graphs (but evenly to ensure similar group sizes).** For the first trial, the program selected one of the four available graphs displaying between-hospitals comparisons. For the second trial, the program selected one of the eight available graphs displaying time-series comparisons. Now we describe the questions asked for each graph.

Between-hospitals comparisons. Participants were asked the following questions about the graphs presenting between-hospitals comparisons; note the first four questions do not require participants to consider chance variation:

(Q1) Which hospital (or hospitals) has the lowest percent compliance?

(Q2) Which hospital (or hospitals) has the highest percent compliance?

(Q3) Which hospital (or hospitals) surveyed the smallest number of patients?

(Q4) Which hospital (or hospitals) surveyed the largest number of patients?

(Q5) Which hospital's (or hospitals') surveyed percent compliance lies beyond what is likely due to chance variation, three standard deviations from the centre line?

Participants responded to each question by selecting the relevant hospital code(s) A-O. Question 5 contained two additional choice responses: None and I can't tell.

---

[4] As we wanted the percentages displayed in the charts to convey something that could actually happen, we followed the same procedure as that described in footnote 3 to flip the run chart data.

(Q6) For this question participants were told to imagine that their GP suspected they might have cancer and would refer them to one of two hospitals with equal performance measures (i.e., either both were 99% or both were 91%). If the graph displayed was a funnel chart, the hospitals were B and M. If the graph displayed was a league table the hospitals were A and B or N and O. Participants were asked to use the information provided in the graph to determine which hospital would most likely see them within two weeks. The four response options included the following: (A) select hospital [insert relevant hospital code], (B) select hospital [insert relevant hospital code], (C) I am equally likely to be seen in two weeks or sooner at either hospital, and (D) I do not understand the question. The correct answer when the graph presented two equally high data points was M for the funnel chart and B for the league table, and the correct answer when the graph presented two equally low data points was B for the funnel chart and O for the league table.

Time-series comparisons. Participants were asked the following questions about the graphs presenting time-series comparisons; note that the first two questions do not require participants to consider chance variation:

(Q7) Which month (or months) has the highest percent of patients waiting over four hours?

(Q8) Which month (or months) has the lowest percent of patients waiting over four hours?

(Q9) Which month's (or months') percent lies beyond what is likely due to chance variation, three standard deviations from the centre line?

Participants responded to each question by selecting the relevant month(s), March - February. Question 8 contained two additional choice responses: None and I can’t tell.

(Q10) For this question participants were told to imagine they were a hospital manager who based on the information provided in the graph needed to decide whether to investigate the change that occurred in February. They were cautioned that while investigations can improve future performance measures, such investigations are costly and they should not investigate data which are likely the result of chance variation. The four response options included the following: (A) Yes, (B) No, (C) I have no preference, and (D) I do not understand this question.

Analyses. To organise our results, we first describe participants' responses to the graphs presenting between-hospitals and then time-series comparisons. We first present descriptive statistics of responses to each of the questions. Next, the participants' responses are dichotomously categorised as either correctly interpreting the information contained in the graph or not (e.g., identifying the statistically designated special-cause datum or not, making the statistically informed decision or not). With these dichotomised responses we use a Chi-square test to compare the performance of those participants who saw a control chart with those who saw a non-control chart. The results for the more difficult questions (i.e., hospital preference or investigative choice) are shown in Figure 2.

Experiment 1. Results

Between-hospitals comparisons.

Similar numbers of participants received each graph type; 81 saw a funnel chart and 86 saw a league table.

Identifying the Most Extreme data points. No matter which chart participants viewed, their responses were largely accurate. Of participants who saw a funnel chart 77% correctly identified the lowest and 70% the highest performing hospital; 85% correctly identified the smallest and 82% the largest sized hospital. Of those who saw a league table 77% correctly

identified the lowest and 84% the highest performing hospital; 93% correctly identified the smallest and 87% the largest sized hospital. The Chi-square tests only found a difference for the question asking participants to identify the highest performing hospital, favouring league tables ($\chi^2(1, N = 170) = 4.37$, $p = 0.04$, $\varphi = 0.16$). None of the other questions yielded statistically different results (ps > 0.08).

Identifying the special-cause datum. Funnel charts increased participants' ability to identify the special-cause datum. Given a funnel chart, 51% correctly identified the special-cause datum; 24% selected one or more incorrect options, and 25% said they could not tell. For the league table only 1% correctly identified the special-cause datum; 48% selected one or more incorrect options, and 51% said they could not tell. A Chi-square test determined that participants presented with a funnel chart responded more accurately than those presented with a league table, $\chi^2(1, N = 170) = 55.44$, $p < 0.001$, $\varphi = 0.57$.

Hospital Choice. Participants struggled to identify the hospital that would most likely see them within two weeks, no matter which graph they saw. Given a funnel chart 43% chose the statistically recommended hospital, 31% chose the other hospital, 24% said they were equal and 2.4% reported that they did not know. For the league table 49% chose the statistically recommended hospital, 27% chose the other hospital, and 24% said they were equal. The Chi-square test did not find a significant difference ($p = 0.43$).

Time-series comparisons.

Similar numbers of participants received each time-series graph type; 87 received a run chart with control lines and 84 a run chart without control lines.

Identifying the Most Extreme data points. No matter which chart participants viewed, their responses were largely accurate. Of participants who saw a run chart with control lines,

88% correctly identified the lowest and 93% the highest performing hospital. Of those who saw a run chart without control lines, 96% correctly identified the lowest and 100% the highest performing hospital. The run chart without control lines encouraged more accurate responses for both these questions as analysed using the Chi-square tests, χ^2s (1, N = 170) > 3.90, ps < 0.05, φ = 0.15.

Identifying the special-cause datum. Run charts with control lines increased participants' ability to identify the presence or absence of the special-cause datum. Of participants who saw a run chart with control lines, 64% correctly identified the presence or absence of the special-cause datum, 16% were incorrect and 20% said they could not tell. Of participants who saw a run chart without control lines, 13% correctly identified the special-cause datum, 36% selected incorrect data and 51% said they could not tell. A Chi-square test found that those participants presented with a run chart with control lines responded more accurately than those presented with a run chart without control lines, $\chi^2(1, N = 170) = 48.7$, $p < 0.005$, $\varphi = 0.54$.

Investigative Choice. When asked whether they would call for an investigation, participants struggled no matter which graph they saw. Of participants who saw a run chart with control lines, 56% chose the statistically recommended course of action, 40% chose the other course of action, 4% had no preference and 1% did not understand the question. Of participants who saw a run chart without control lines, 48% chose the statistically recommended course of action, 45% chose the other course of action, 6% had no preference, and 1% indicated they did not understand the question. A Chi-square test did not find a significant difference, (p = 0.29).

Experiment 1. Discussion

This experiment compared lay-people's use of control charts to non-control charts. Confirming our hypotheses, control charts helped participants identify the special-cause datum;

but did not help participants answer more difficult questions (i.e., hospital preference or investigative choice). This suggests training people to use control charts should not focus on helping them identify special-cause data; rather training should focus on how people should apply such information to support more difficult decisions.

Regarding hospital choice, one may note that most people do not consider quantitative healthcare data when choosing which hospital to attend. Rather people are more likely to rely on their GP's advice or the experiences of their friends (Department of Health 2009; Dixon, et al., 2010). However, this should not be taken to mean people do not want more reliable information. People report that health websites allowing them to compare hospitals are a useful source of information, which they believe should be more widely available (Boyce, et al., 2010). To support people's desire for such information, general practitioners could inform them that such websites are available when patients are most likely to want this information. Our general discussion provides an idea for how to make the information provided in control charts easier to interpret thus allowing people to better account for the role of chance in their decisions.

The next section discusses Experiment 2. This experiment was conducted to account for two concerns presented by Experiment 1 about the results for graphs presenting time-series comparisons. The first is that our participants were inexperienced with investigative **hospital** questions and so our results may not generalise to **hospital decision-makers**. The second concern is that participants in Experiment 1 were only asked whether they would call for an investigation based on the data provided. Such a question conflates participants' ability to interpret the statistical recommendations contained in the graphs with the desire to actually investigate. However, as lay-people have no experience with actually calling for an investigation there may be no meaningful difference between these questions for them. In contrast, there is

more likely a difference between the questions for experienced hospital decision-makers. To address these concerns, in Experiment 2 we surveyed hospital decision-makers who are experienced in these choices, and replaced the investigative choice question with questions about the graphs' investigative recommendations and their actual investigative choices.

Experiment 2. Hospital Decision-Makers

Methods

Participants. A lead consultant invited high level hospital decision-makers within one United Kingdom NHS trust to voluntarily complete this survey. In total, 47 participants (45% female) completed the survey (2.1% under 31 years old; 68.1%= 31-50 years old; 29.8% 51-70 years old). These participants included consultants (60%), managers (25%), doctors (9%), nurses (4%), and 1 clinical audit adviser (2%). Most (87%) had two or more years of work experience in a decision-making capacity. Nearly three quarters of the participants recalled completing a formal statistics course (70%) and more than half knew of control charts (60%). Following completion of the survey, participants were entered into a lottery draw for a chance to win an Amazon gift voucher.

Materials. The survey was created and delivered on-line using *Qualtrics*©2012. The time-series graphs of Experiment 1 were used but with one alteration. This alteration was to add the sample-size aside each month on the horizontal axis. Although sample-size information is rarely included in charts, such information affects where the control lines are set, which could plausibly be considered useful, and so we wanted it to be available for the hospital decision-makers.

Procedure. Participants were presented with two graphs presenting time-series comparisons. The same questions used in Experiment 1 were used to assess participants' ability to interpret the graphs and identify the presence or absence of special-cause data (Q7-Q9). To address the concerns raised for Experiment 1, the previously used investigative choice question was replaced with two questions designed to distinguish participants' abilities to interpret the graphs' investigative recommendations and their actual investigative choices.

(Q11) The question requiring participants to interpret the graphs' investigative recommendations asked: "Regardless of what you would actually do, does the information provided in the chart above suggest you should call for an investigation to learn more about the increase (or decrease) in February?" The four response options included the following: (A) Yes, (B) No, (C) This chart makes no recommendations about what I should do and (D) I do not understand this question.

(Q12) The question asking for participants' actual investigative choices asked: "In reality, if these performance data were from your hospital would you call for an investigation to learn more about the increase (or decrease) in February?" The four response options included the following: (A) Yes, (B) No, (C) I don't know and (D) I do not understand this question. After responding, participants were asked to further explain why they would or would not actually investigate based on data like those presented in the graph.

Experiment 2. Results

The results are reported in a similar manner as Experiment 1. However, due to the small expected values in Experiment 2 (a result of a smaller sample size), **Fisher's** exact test was applied rather than the Chi-squared test. Similar numbers of participants received each time-

series graph type; 24 received a run chart with control lines and 23 received a run chart without control lines.

Identifying the Most Extreme data points. No matter which chart participants viewed, their responses were largely accurate. Of participants who saw a run chart with control lines, 100% correctly identified the lowest and 100% the highest performing hospital. Of those who saw a run chart without control lines, 96% correctly identified the lowest and 91% the highest performing hospital. Neither question statistically differed when analysed using Fisher's test (*p*s > 0.23).

Identifying the special-cause datum. The control charts increased participants' ability to identify the presence or absence of the special-cause datum. Of participants who saw a run chart with control lines, 63% correctly identified the presence or absence of the special-cause datum, 13% were incorrect and 25% said they could not tell. Of participants who saw a run chart without control lines, only 4% correctly identified the special-cause datum, 17% selected incorrect data and 78% said they could not tell. Fischer's test showed that those presented with a run chart with control lines responded more accurately than those presented with a run chart without control lines ($p < 0.001$).

Investigative recommendation. The statistical recommendations given by control charts were accurately interpreted by most participants. Of the participants who saw a run chart with control lines, 79% correctly identified the course of action the control chart recommended, 8% chose the other course of action, 13% said the chart made no recommendations. Of the participants who saw a run chart without control lines, 9% chose the recommended course of action, 26% choosing the other course of action, 61% said the chart made no recommendations and 4% did not understand the question. A Fischer's test showed that those presented with a run

chart with control lines responded more accurately than those presented with a run chart without control lines ($p < 0.001$).

Investigative choice. The recommendations made by the control charts were largely followed by our participants. Of the participants who saw a run chart with control lines, 68% said they would follow the statistically recommended course of action, 29% chose the other course of action, 4% did not know. Participants who saw a run chart without control lines were more evenly split; 39% would follow the statistically recommended course of action, 22% chose the other course of action, 40% said they did not know. A Fisher's test showed that those presented with a run chart with control lines responded more accurately than those presented with a run chart without control lines ($p < 0.001$).

Experiment 2. Discussion

Experiment 2 compared hospital decision-makers' use of run charts with control lines to run charts without control lines. Notably, when provided with a run chart with control lines, as opposed to one without, hospital decision-makers' were better able to identify the special-cause datum, interpret the recommendations made by the chart and apply those recommendations to their investigative choice. These results support our first, but not second hypothesis.

General Discussion

Control charts facilitate statistically informed decision-making better than non-control charts. This is true for both lay-people and hospital decision-makers. The advantage of using control charts is more easily observed with simple questions (is there any special-cause data?), than more difficult questions (which hospital is more likely to see you within two weeks, or would you call for an investigation?). In the following discussion we first compare and contrast

the responses of lay-people and hospital decision-makers. Lastly, we close with recommendations to help people effectively employ control charts.

Limitations

It should be noted that our lay-people were recruited from a university and many had statistical training. While such people are part of the general population, further work with non-university samples would be a welcomed addition to this literature. **Our sample of lay-people was a convenient sample with which we could initially explore how control charts influence people's decisions**.

Another limitation of the current study is that we did not assess hospital decision-makers use of funnel charts and league tables, and so cannot compare lay-people and hospital decision-makers' use of them. We did not assess hospital decision-makers use of funnel charts and league tables for at least two reasons. First, a randomized controlled trial of this comparison was already performed in 2004 (Marshall et al.). Second, the hospital requested we keep the survey brief.

Comparing lay-people and hospital decision-makers

Descriptively our lay-people and hospital decision-makers exhibited some similarities in their choices. For Q7, when using a control chart 63% of hospital decision-makers and 64% lay-people correctly identified the presence or absence of the special-cause datum. Hospital decision-makers were more cautious. Provided with a run chart without control lines, approximately three in every four hospital decision-makers said they could not tell whether any data were special-cause, fewer (approximately two in every four) lay-people did. This difference may reflect hospital decision-makers being better acquainted with the costs of making a mistake, and their need to be more certain before answering statistical questions. Control charts provide such certainty.

A major difference between populations was their ability to use the control lines to decide whether to investigate. Lay-people's investigative choices were not affected by the control lines, but hospital decision-makers' choices were. This difference may suggest that lay-people do not understand or appreciate the thinking behind decisions that policy makers and hospital decision-makers make. For example, they may not understand how costly investigations may become, how often data appear aberrant or how systems changes need not include investigations of particular data.

The current study shows that control charts are indeed useful. Hospital decision-makers should use them more often. While the use of control charts in healthcare is increasing, they are still very much underused (Koetsier, et al., 2012; Taylor, et al., 2014). In organizations where control charts are not presently used, a prevailing social norm may prevent their introduction (For a demonstration of control charts being introduced into a plausibly resistant setting, see de Leval et al., 1994). Our work demonstrates empirically some benefits that control charts offer and therefore may provide the spark that organizations need to start using them.

Helping people effectively employ control charts.

People may have problems interpreting graphs in general. To help these people, Rakow, et al., 2014 note that carefully constructing the graphs so that the axes provide the right information (**e.g.,** mortality vs survival) is an important step to take. Control charts however offer an additional challenge, in that if people do not know what control lines represent they will likely ignore them. Indeed, Zikmund-Fisher's (2007) research found that people are largely unfamiliar with control charts and experience difficulty interpreting them. This is not a problem with control charts, but rather the users' ability to interpret them. This barrier may be overcome with educational interventions. **Curran, et al. (2008) used a very light** educational intervention,

wherein their control charts were accompanied by one or two sentences describing the observed variation and any out-of-control episodes or trends.

Statistically more astute decision-makers may also experience difficulties using control charts. Wheeler (2011) and Balestracci (2011) offer more complete discussions of such concerns. Based on their work, we provide a brief description of two of the concerns below.

Concern 1. Many statistically **astute** people may think that if control charts data are not normally distributed then you cannot use a control chart. This premise is wrong. The control charts' use is largely insensitive to the data's distribution, which is particularly true for individual control charts (Wheeler, 1995). When organizations measure a process over time, they often do not know whether the process is stable enough for the data to be treated as if it had come from a single population (with a well-defined distribution); control charts help determine whether this is the case. Even if the data to be plotted are not normally distributed, control charts still have a reasonable false positive rate and any rule breaks warn that the data may be coming from different processes.

For quality improvement efforts in hospital organisations, a further distinction should be made between two phases of control chart use. In Phase one control charts are used in an explorative and iterative fashion to eradicate factors that cause worrisome variations and in so doing create an in-control or stable process. Ensuring process stability in itself often improves quality performance, but such improvements may not reach the desired specifications. Phase two can then be used to promote further quality improvements. Similar to hypothesis testing, in Phase two control charts are used to detect deviations from an expected distribution (aka outliers in the hypothesis testing literature). Phase one is a necessary step before advancing to Phase two

because if existing special-causes are not yet understood, they will complicate experimental attempts to improve the process (for more information see: Woodall, 2000).

Concern 2. Another concern raised by the statistical astute regards where the control lines are set. Determining, precisely where the control lines should be set on charts for different measures should reflect the cost of investigation and the cost of not investigating, in terms of financial costs, quality, and harm. This is a question of judgement and cannot be resolved statistically, i.e., that the lines on a control charts should always be set a number of deviations from the central tendency. When more variation is acceptable, then wider set control lines **(lower sensitivity)** may be appropriate. **However the precise position of the control lines must arise from the data themselves, as opposed to the precise position of a target lines which may be based on a precise external standard (e.g., 95% of patients attending an Accident and Emergency department are seen within four hours of arrival).** Additionally, the control lines need not reflect a normal distribution; for example, a binomial distribution is often used to set control limits for proportion data.

To briefly demonstrate a consequence of setting the control lines more or less conservatively the following case is offered. Given a normal distribution, setting the control lines at three standard deviations is very conservative. At three standard deviations, the control lines will encompass about 99.5% of the data in a stable process, creating a small false alarm rate 0.25%. But here the astute statistician notes that such a conservative setting also increases the number of times truly concerning data is overlooked, i.e., the miss rate. If indeed the distribution is normal one may be more comfortable setting the control lines at two standard deviations, so that about 95% of the data in a stable process are encompassed, the false alarm rate is still only 5%, and the number of misses is reduced.

However, data are not always normally distributed. As stated in the introduction, if the data come from an arbitrary distribution, 25% of data can be located beyond two standard deviations depending on the shape of the underlying distribution (Chebyshev's inequality), and so the chance for false alarms is quite high (Kvanli, Pavur & Keeling, 2006). Put another way, if hospital decision-makers set the control limit at two standard deviations, this could prompt them to investigate up to 1 in every 4 data; in most cases this would be infeasible. In contrast, if the control lines are set at three standard deviations, this could prompt them to investigate 1 in every 10 data. Thus setting the control lines at three standard deviations may not be as conservative as first thought.

Ultimately, no probabilistic detector (including intuition) is 100% accurate, but investigating every data point is infeasible. Compared to intuitions, control charts allow one to better understand variation in performance measures and in so doing better focus quality improvement efforts. Without control lines variation is often ignored, so at the very least control lines prompt crucial discussions about the costs and benefits of investigative action. Specifically, when data are statistically irregular, they are more likely to be the result of a special-cause that an investigation can identify to guide quality improvement efforts. When data are statistically regular, i.e., common-cause data, there is nothing an investigation can find and so investigations absorb resources that could be better spent elsewhere. Where common-cause data are unacceptable quality improvement efforts should focus on modifying processes that are common to all the data.

References

Alemi, F., & Neuhauser, D. (2004). Time-between control charts for monitoring asthma attacks. *Joint Commission Journal on Quality Improvement and Patient Safety, 30*, 95-102.

Amin, S. G. (2001). Control charts 101: A guide to health care applications. *Quality Management in Health Care, 9*(3), 1-27.

Balestracci, D. (2011). Four control chart myths from foolish experts: Don't teach people statistics—teach them to solve problems. Quality Digest, Retrieved March 23, 2015, from: http://www.qualitydigest.com/inside/quality-insider-article/four-control-chart-myths-foolish-experts.html

Boyce, T., Dixon, A., Fasolo, B., et al. (2010). Choosing a high quality hospital: the role of nudges, scorecard design and information. The Kings Fund: London.

Carey, R. G., & Teeters, J. L. (1995). CQI case-study—reducing medication errors. *Joint Commission Journal on Quality Improvement, 21*, 232-237.

Chance, B. L. (2002). Components of Statistical Thinking and Implications for Instruction and Assessment, *Journal of Statistics Education [Online], 10*(3) www.amstat.org/publications/jse/v10n3/chance.html.

Curran, E. T., Benneyan, J. C., & Hood, J. (2002). Controlling Methicillin-Resistant Staphylococcus Aureus: A feedback approach using annotated statistical process control charts. *Infection Control and Hospital Epidemiology, 23*, 13-18.

Curran, E., Harper, P., Loveday, H., Gilmour, H., Jones, S., Benneyan, J., Hood, J., & Pratt, R. (2008). Results of a multicentre randomised controlled trial of statistical process control charts and structured diagnostic tools to reduce ward-acquired meticillin-resistant

Staphylococcus aureus: the CHART Project. *Journal of Hospital Infection, 70*(2), 127-35. doi: 10.1016/j.jhin.2008.06.013

de Leval, M.R., Francois, K., Bull, C., Brawn, W. & Spiegelhalter, D. (1994). Analysis of a cluster of surgical failures: application to a series of neonatal arterial switch operations. *Journal of Thoracic and Cardiovascular Surgery, 107*, 914-924.

Deming, W. E. (1975). On probability as a basis for action, *The American Statistician*, *29*(4), 146-152. doi: 10.1080/00031305.1975.10477402

Department of Health. (2009). Report on the National Patient Choice Survey – March 2009 England [online]. Available at: ww.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_103681.pdf (accessed on 27 September 2015).

Dixon, A., Robertson, R., Appleby, J., Burge, P., Devlin, N., Magee, H. (2010). Patient choice: How patients choose and how providers respond. London: The King's Fund. Available at: www.kingsfund.org.uk/publications/patient_choice.html. (accessed on 28 September 2015).

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A*, 159, 385-443.

Hildon, Z., Allwood, D., & Black, N. (2012). Making data more meaningful: Patients' views of the format and content of quality indicators comparing health care providers. *Patient Education and Counseling, 88*(2), 298-304. doi: 10.1016/j.pec.2012.02.006

Kahneman, D. (2003). A perspective on judgement and choice. *American Psychologist, 58*, 697-

720. doi:10.1037/0003-066x.58.9.697

Koetsier, A., van der Veer, S. N, Jager, K. J., Peek, N., & de Keizer, N. F. (2012). Control charts in healthcare quality improvement: A systematic review on adherence to methodological criteria. *Methods of Information in Medicine. 51*(3), 189-198.

Kvanli, A. H., Pavur, R. J., & Keeling, K. B. (2006). *Concise Managerial Statistics*. *cEngage Learning.* 81-82. ISBN 9780324223880

Marshall T, Mohammed MA, & Rouse A. A (2004). A randomized controlled trial of league tables and control charts as aids to health service decision-making. *International Journal for Quality in Health Care, 16*(4), 309-315.

Mohammed, M., Worthington, P., & Woodall, W. H. (2008). Plotting basic control charts: tutorial notes for healthcare practitioners. *Quality and Safety in Health Care, 17,* 137-145. doi: 10.1136/qshc.2004.012047

Perla, R. J., Provost, L. P., & Murray, S. K. (2011). The run chart: a simple analytical tool for learning from variation in healthcare processes. *BMJ Quality and Safety, 20*(1), 46-51. doi: 10.1136/bmjqs.2009.037895

Peymané, A., Rouse, A. M, Mohammed, M. A, Marshall T. (2002). Performance league tables: The NHS deserves better. BMJ, 324(7329), 95–98.

Polit, D. F. & Chaboyer, W. (2012). Statistical process control in nursing research. *Research in Nursing and Health, 35,* 82-93.

Rakow, T., Wright, R. J., Spiegelhalter, D. J., & Bull, C. (2014). The pros and cons of funnel plots as an aid to risk communication and patient decision making. *British Journal of Psychology, 106*(2), 327-348.

Shewhart, W. A. (1939). *Statistical Method from the viewpoint of Quality Control*. Graduate

School of the Department of Agriculture, Washington, D.C.

Speekenbrink, M., Twyman, M., Harvey, N. (2012). Change detection under autocorrelation. CogSci 2012 Sapporo, Japan. Proceedings of the 34th Annual Conference of the Cognitive Science Society.

Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine, 24*(8), 1185-1202. doi: 10.1002/sim.1970

Sunstein, C. R. (2005). Moral Heuristics. *Behavioral and Brain Sciences*, *28*(4), 531-542. doi: 10.1017/S0140525X05000099

Taylor, M. J., McNicholas, C., Nicolay, C., Darzi, A., Bell, D., & Reed, J. E. (2014). Systematic review of the application. BMJ Quality & Safety, 23(4), 290-298. doi: 10.1136/bmjqs-2013-001862.

Thor, J., Lundberg, J., Ask, J., Olsson, J., Carli, C., Harenstam, K. P., & Brommels, M. (2007). Application of statistical process control in healthcare improvement: Systematic review. *Quality Safety in Health Care*, *16*, 387-399. doi: 10.1136/qshc.2006.022194.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Wainer, H. (2013). Medical Illuminations: Using Evidence, Visualization and Statistical Thinking to Improve Healthcare. Oxford University Press. ISBN 0199668795.

Wheeler, D. J. (1995). *Advanced Topics in Statistical Process Control*. Statistical Process Controls, Inc., Knoxville, TN.

Wheeler, D. J. (2011). Myths about process behavior charts: How to avoid some common

obstacles to good practice. *Quality Digest*, retrieved March 23, 2015, from: http://www.qualitydigest.com/inside/quality-insider-article/myths-about-process-behavior-charts.html#

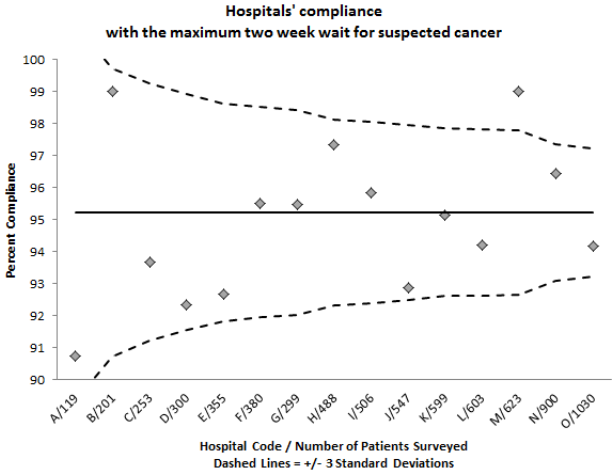
Woodall, W. H. (2000). Controversies and contradictions in statistical process control. *Journal of Quality Technology, 32*(4), 341-350.

Woodall WH. (2006). The use of control charts in health-care and public-health surveillance (with discussion). *Journal of Quality Technology, 38*, 104-105.

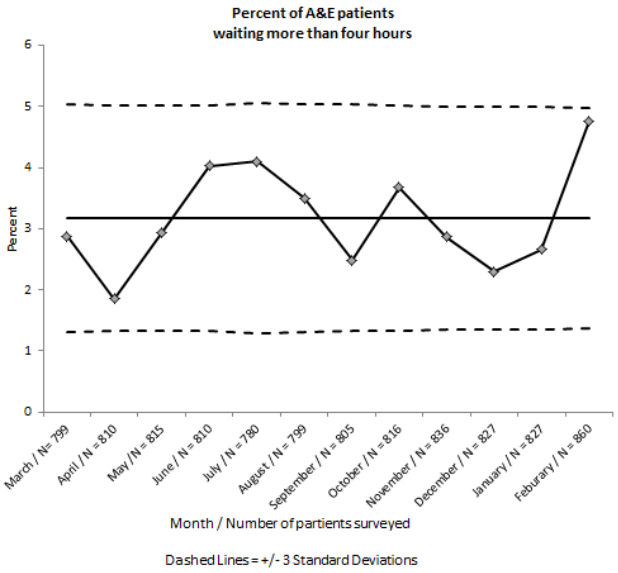
Zikmund-Fisher, B. J., Smith, D. M , Ubel, P. A., & Fagerlin, A. (2007). Validation of the subjective numeracy scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making, 27*, 663-671.

Control charts

A. Funnel chart

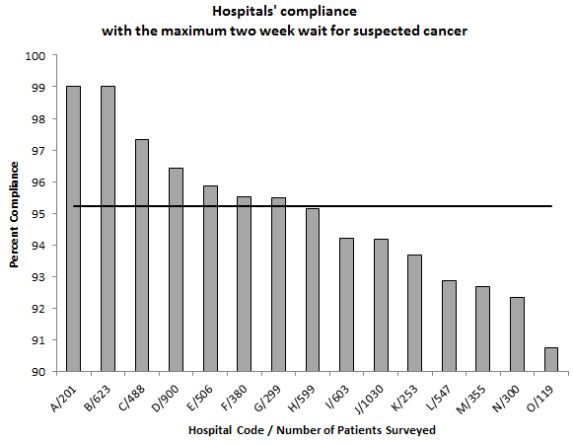


B. Run chart with control lines



Non-control charts

C. League table



D. Run chart without control lines

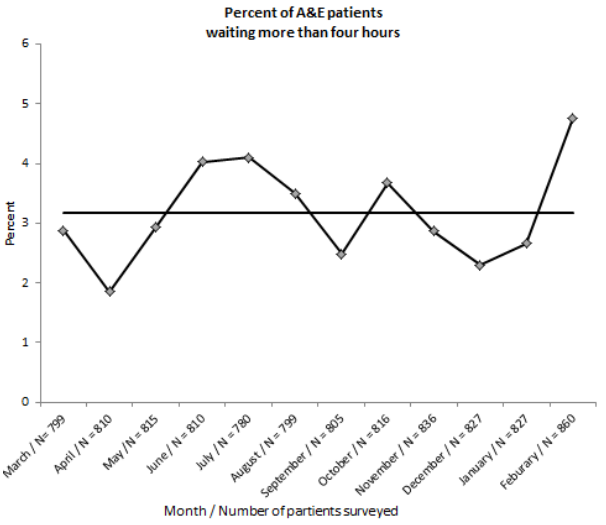


Figure 1. Example control charts and non-control charts presenting between-groups (funnel charts and league tables) and time-series comparisons (run charts).

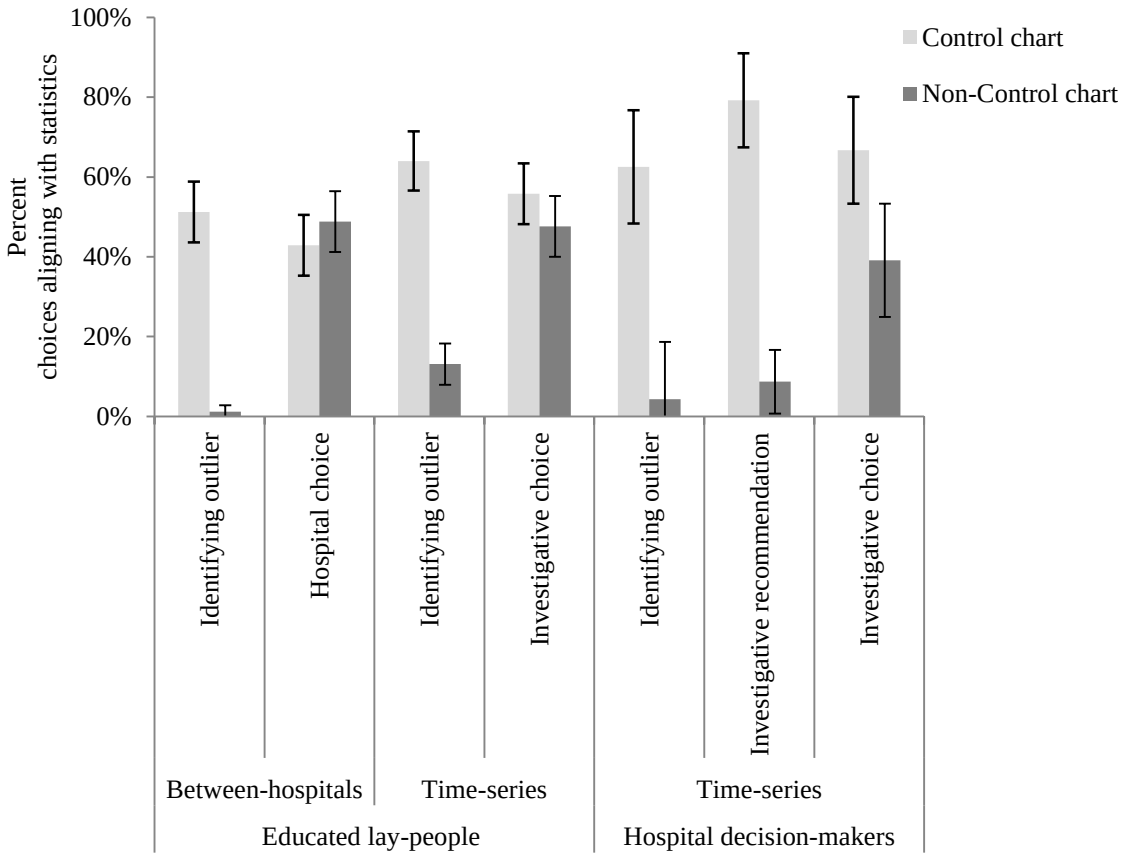


Figure 2. A graphical representation of the Chi-square tests. Error bars represent 2 SEs.

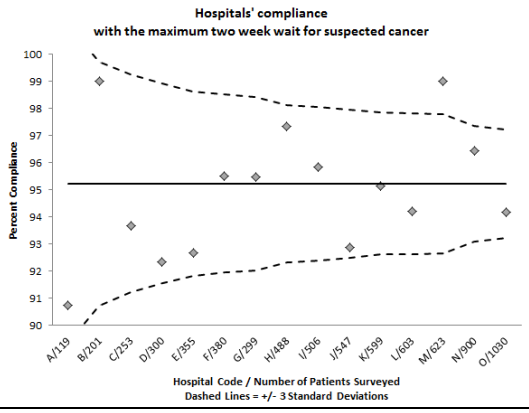
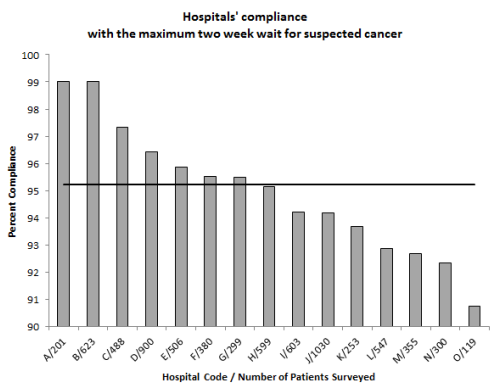
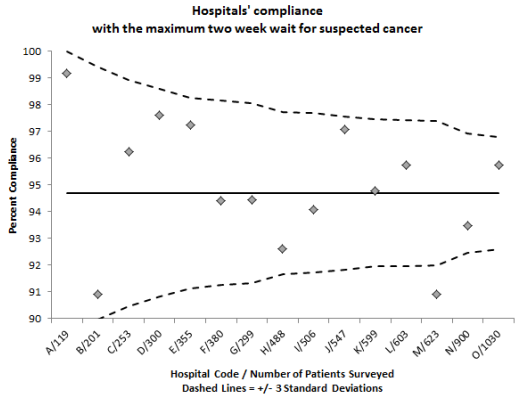
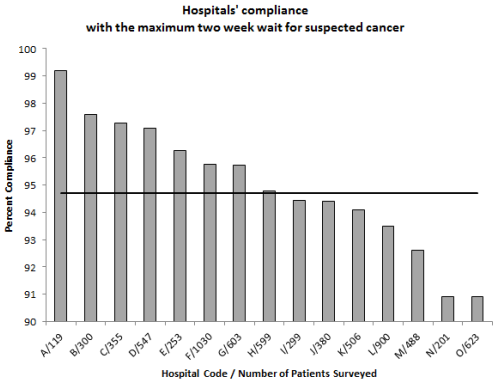
658 Figure Captions

659 *Figure 1*. Example control charts and non-control charts presenting between-groups (funnel
660 charts and league tables) and time-series comparisons (run charts).

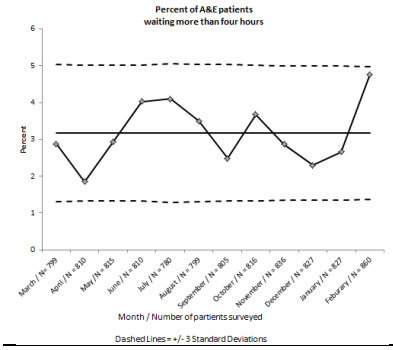
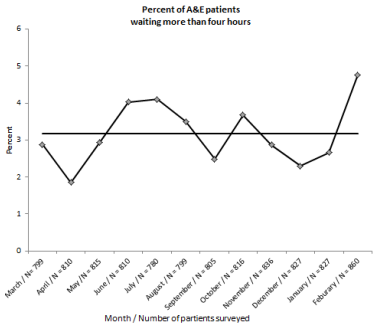
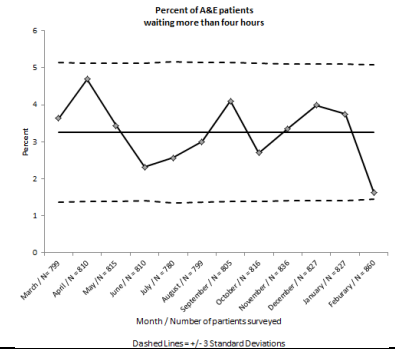
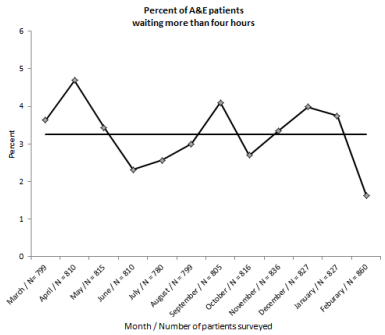
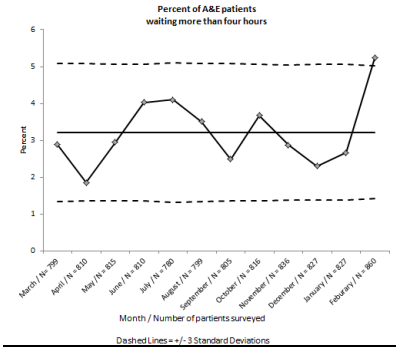
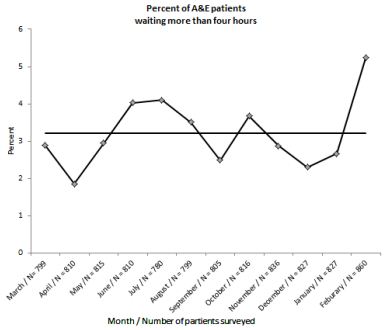
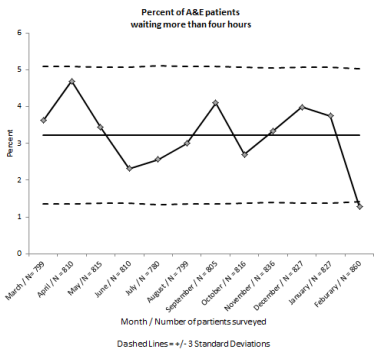
661 *Figure 2*. A graphical representation of the Chi-square tests. Error bars represent 2 SEs.

662

Appendix A. Graphs presenting between subjects comparisons.

	Funnel Chart	League Table
Data set with two equally high (potentially outlying) data points	A 	B 
Data set with two equally low (potentially outlying) data points	C 	D 

Appendix B. Graphs presenting time-series comparisons.

	Run Chart with control lines	Run Chart without control lines
In control Increasing	A 	B 
In control Decreasing	C 	D 
Out of control Increasing	E 	F 
Out of control Decreasing	G 	H